Data Reduction Technique for Large Streaming Data by Locally Exchangeable Measures

Scientific Data Management Group
Computational Research Division
Lawrence Berkeley National Laboratory

In collaboration with ESnet

Background

- **Performance analysis/modeling**
  - Next generation analytic computing platform/environment
    - Studies on data access patterns, data I/O issues, system performance, etc
  - Network performance analysis/modeling with monitoring data
- **Observations**
  - Large streaming data needs a big storage.
  - Exact compression on big streaming data is intractable, in general.
    - Typical alternative: random sampling
      - It is not scalable for high-rate multiple streaming data
      - There is no guarantee of reflecting the underlying data distribution
  - Large streaming data tend to show redundant data patterns.
  - Statistical analysis is needed on big data.
  - Many conventional statistical methods are based on a specific assumption (exchangeability).

Exchangeable Random Variables

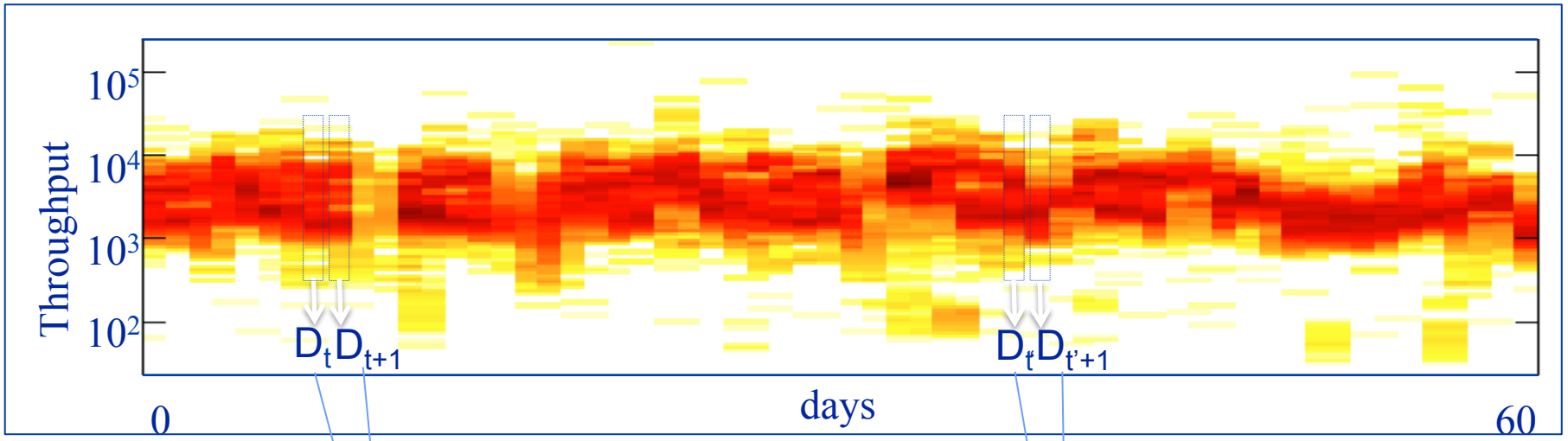
Exchangeable RVs: a set of RVs which are interchangeable among others.

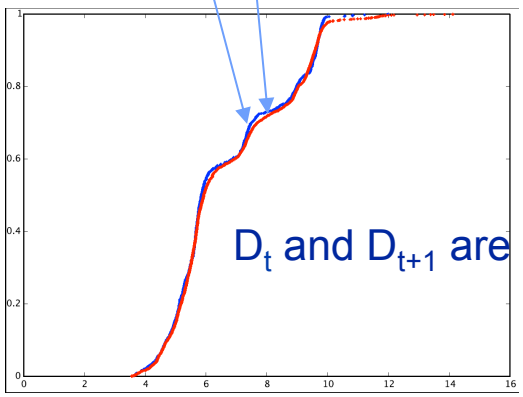
$$P(x_1, \cdots, x_n) = P\left(x_{\pi(1)}, \cdots, x_{\pi(n)}\right)$$ π: a permutation

- **Exchangeability is already exploited and utilized in many applications such as image and video retrieval.**
- **Examples**
  - **Image & video matching: exchangeable image features**
  - **Econometrics: a set of exchangeable portfolio (in risk analysis)**

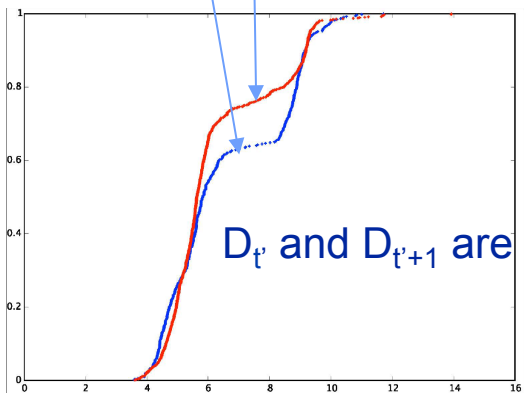
An example: network monitoring data

- **Checking exchangeable blocks by building cumulative histograms**





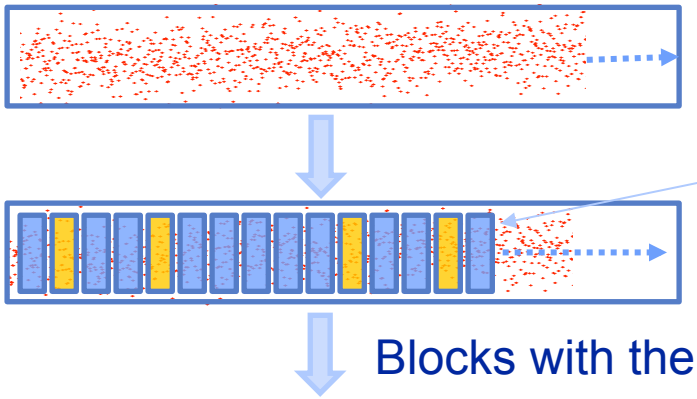
D_t and D_{t+1} are exchangeable



$D_{t'}$ and $D_{t'+1}$ are not exchangeable

An Example of Locally Exchangeable Measures (LEMs)

Input: streaming data



Blocks with the same color present LEMs.

Data Reduction Algorithm by
Locally Exchangeable Measures

Output:

Sampled Data for Statistical Analysis

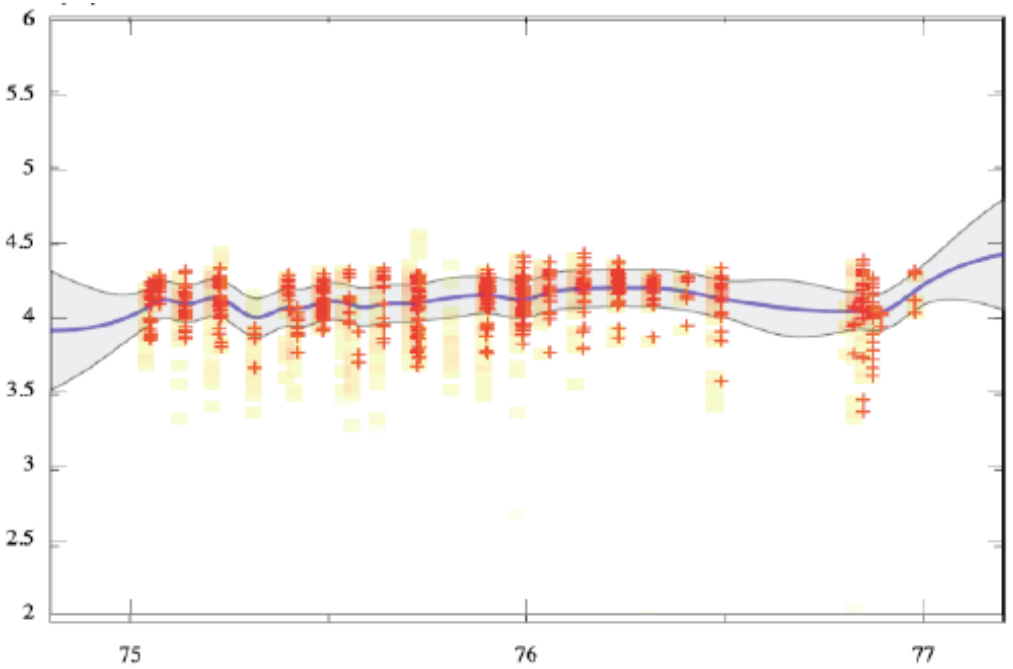
Experimental results on network monitoring data

- **Data reduction rates**

Router	Total records	Sampled records	Reduction Rate
RT1	33.6 million	11.3 million	66.5%
RT2	28.1 million	14.7 million	47.5%
RT3	15.8 million	3.0 million	80.9%
RT4	14.4 million	2.6 million	71.6%
RT5	9.2 million	2.7 million	70.9%
RT6	10.8 million	2.9 million	73.6%
Total	112.5 million	37.8 million	66.4%

Application on Gaussian Process

- Gaussian Process (GP) is a popular regression method for streaming data.
- Computational complexity in GP is $O(n^3)$.
  - Thus, conventional methods do not scale.
- LEMs can be used to generate accurate samples for efficient GP.



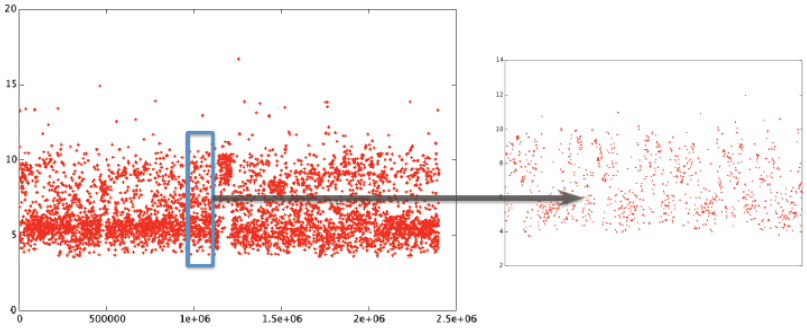
Summary

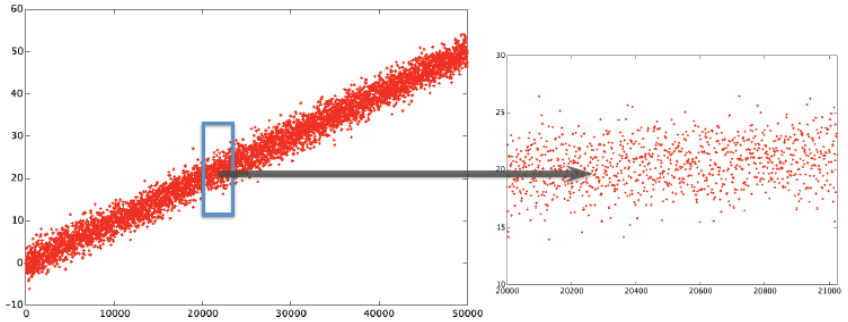
- **Statistical analysis enables estimating future events in various applications. For example,**
  - Financial market analysis
  - Environmental study
  - Energy usage analysis
  - Social network media analysis
  - Traffic analysis
  - System performance monitoring analysis
- **Locally Exchangeable Measures (LEMs)**
  - Enables efficient data reduction on the large streaming data
  - Provides accurate statistical analysis without loosing the underlying data distribution
  - Can be applicable to large data archives (offline data)
    - for pattern searching and data reduction
  - An Efficient Data Reduction Method with Locally Exchangeable Measures
    - U.S. Provisional Patent Application, serial no. 61/909,518, filed 11/27/2013.
- **Questions? contact Alex Sim <ASim@LBL.Gov>**

Backup slides

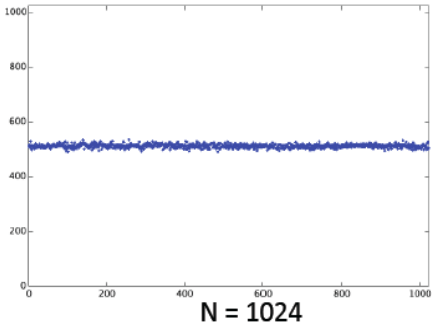
Example for exchangeability test

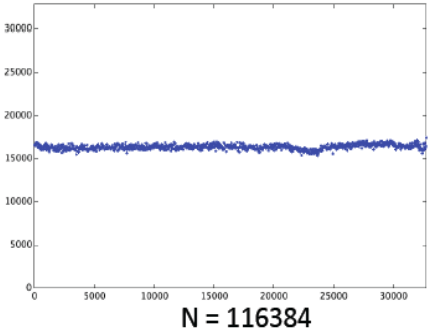
- **Network monitoring data is locally exchangeable**



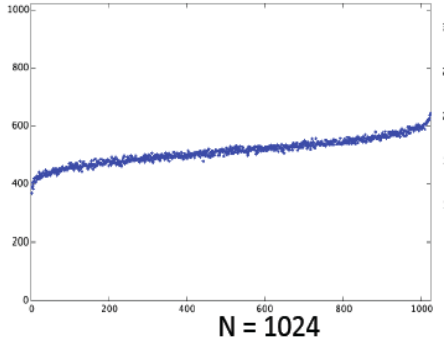


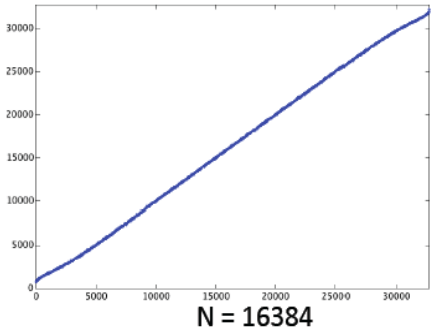
Network monitoring data is exchangeable, when the sampling rate N is small as well as large.





Simulated linear Gaussian is exchangeable, when the sampling rate N is small, but not exchangeable when N is large.





Kolmogorov-Smirnov test (K-S test)

- Statistical hypothesis testing by K-S test to check exchangeable blocks
  - $KS(D_t, D_{t+1}) = \max_l(|F_{D_t}(l) - F_{D_{t+1}}(l)|)$

KS score

  - $F_D(l) = \frac{1}{N}\sum_{\substack{x_i \in X \text{ s.t. } 1\{x_i \le l\} \\ 1 \le i \le |D|}}$

Empirical Cumulative Density Function (ECDF)